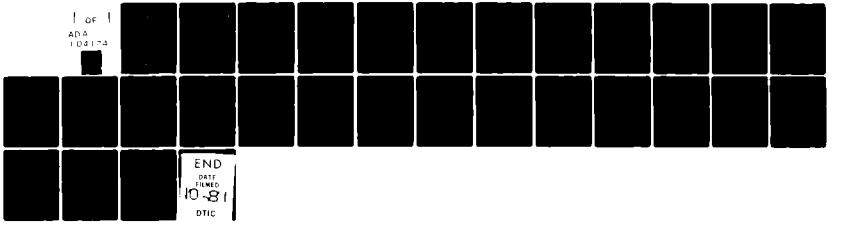
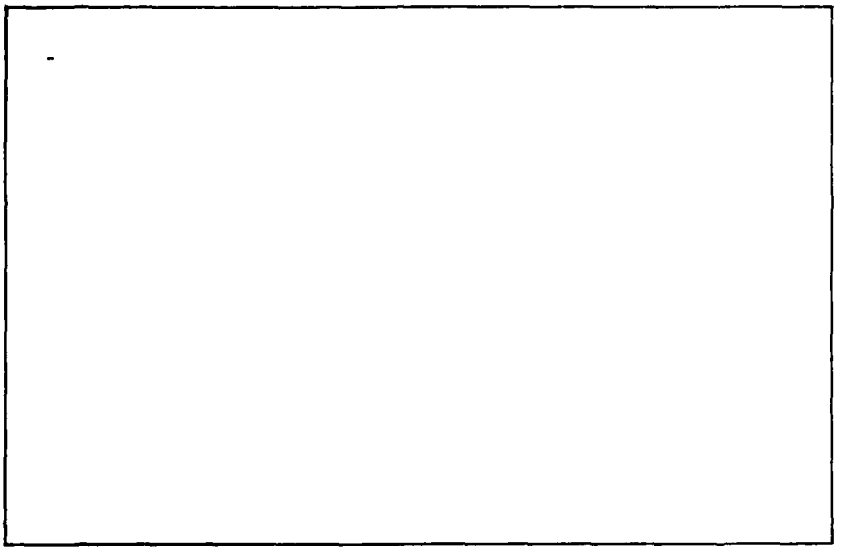
AD-A104 174 CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS F/G 12/1
ASSESSING PROBABILITY ASSESSORS: CALIBRATION AND REFINEMENT.(U)
JUL 81 M H DEGROOT, S E FIENBERG N00014-80-C-0637
UNCLASSIFIED TR-205 NL



LEVEL II

(1)



DEPARTMENT

OF

STATISTICS

Carnegie-Mellon University

PITTSBURGH, PENNSYLVANIA 15213

Assessing Probability Assessors: Calibration and Refinement

by

Morris H. DeGroot
and
Stephen E. Fienberg

Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213

Technical Report No. 205

May, 1981

Revised July, 1981

The preparation of this paper was partially supported by the National Science Foundation under grant SES-7906386 and by the Office of Naval Research Contract N0014-80-C-0637 at Carnegie-Mellon University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

1. Introduction

You have just been hired by the management of a local television station to assist them in evaluating the candidates for a soon-to-be-filled position as station weatherman. Each of the candidates has made a sequence of probability forecasts of the event "rain", announcing the probability p_j on the j^{th} trial of the sequence. Before making the next forecast the candidate learns the value of y_j, which is 1 if "rain" occurs, and is 0 otherwise. The basic data available to you for each candidate is a set of pairs $\{(p_j,y_j): j=1,2,\ldots, n\}$, and from this information you are to assess the candidates, and possibly determine which is the best probability assessor. The purpose of this paper is to provide a probabilistic framework into which to set this problem of assessing probability assessors.

In the weatherman problem, we have taken care to ensure that each forecast is made in light of full information of the outcome of previous forecasts, i.e., with feedback. From a subjective probability perspective the announced probability forecasts form a sequence of conditional probabilities in which each term expresses the candidate's degree of belief given all of the information available at the time of the forecast. The probability distribution of these conditional probabilities, found by letting the number of trials $n\to\infty$, is of central concern in this paper.

The notion of calibration concerns the relationship between the probability distribution of conditional probabilities and the long-run frequencies of rain given a particular probabilistic assessment value. Roughly speaking a probability assessor is said to be well-calibrated if, for those trials on which he forecasts the probability x, the long-run frequency of rain is x. Pratt (1962) and Dawid (1981) show that a probability assessor who is coherent in the sense

of de Finetti (1937) must be well-calibrated almost surely. In Section 2, we make more formal this notion of calibration, and, in Section 3, we show that some well-calibrated forecasters are clearly superior to others. We suggest a formal sense in which a given well-calibrated forecaster can be "more refined" and thus "better" than another. Then in Section 4, we demonstrate the link between the concept of refinement and that of sufficiency in the comparison of experiments. This link leads in Section 5 to a rather simple condition for determining whether one well-calibrated forecaster is more refined than another. In Section 6, this condition is exploited in order to determine a "least-refined" forecaster.

Calibration and refinement, as presented in this paper, refer only to the full probability distribution of the assessor's conditional forecasts. However, in the television station example which began this section, and elsewhere in statistical practice, we do not know either this distribution or the long-run frequencies of rain. All that we get to see is a finite set of forecasts and the associated indicators of whether or not rain occurred, i.e., $\{(p_j, y_j): j = 1,2,\ldots, n\}$. In Section 7, we briefly review some scoring rules suggested for such sample situations, and indicate how they relate to the probabilistic concepts of calibration and refinement.

For the forecasting problems considered in the first six sections of the paper there are only two possible outcomes, rain or no rain. We take care in these sections to preserve the orientation of outcomes and work only with the forecasters' assessments of the probability of rain. Kadane and Lichtenstein (1981) show that the loss of orientation leads to the inability to recalibrate a forecaster's assessments. Finally, in Section 8 we discuss extensions of the calibration and refinement structure to forecasting

problems with $s>2$ outcomes. In these problems, we require the ordered vector of assessed probabilities for the s possible outcomes and the associated indicator vector which summarizes which outcome actually occurs.

2. Well-calibrated forecasters

Consider a weather forecaster who day after day must specify his subjective probability x that there will be at least a certain amount of rain at some given location. For simplicity, we shall refer to the occurrence of this well-specified event as "rain." Thus, we may say simply that, at the beginning of each day, the forecaster must specify his probability of rain, and that at the end of each day it is observed whether or not rain actually did occur.

We shall refer to the probability x specified by the forecaster on any particular day as his prediction. Both for realism and simplicity, we assume that the prediction x is restricted to a finite set of values $0=x_0<x_1<\ldots<x_k=1$. (In many weather forecasts, $k=10$ and $x_j=j/10$.) We assume that the forecaster's predictions can be observed over a large number of days, and we shall let $\nu(x)$ denote the probability function (or frequency function) of his predictions over those days. Thus, we can think of $\nu(x)$ either as the probability that the forecaster's prediction on a randomly chosen day will be x, or in the frequency sense as the proportion of days on which his prediction is x. We shall let $\mathcal{X}$ denote the set of possible predictions $\{x_0,x_1,\ldots,x_k\}$ and let $\mathcal{X}^+$ denote the subset of $\mathcal{X}$ containing only those points for which $\nu(x)>0$.

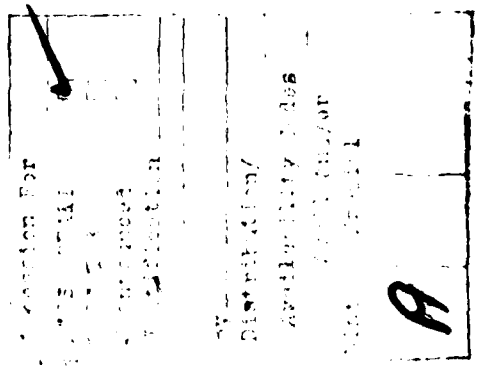
To evaluate the forecaster, we must compare the actual occurrences of rain or no rain with his predictions, and for $x\in\mathcal{X}^+$ we shall let $\rho(x)$ denote the conditional probability of rain given that the prediction is x. The

forecaster is said to be well-calibrated (see, e.g., Dawid, 1981) if $\rho(x) = x$ for all values of $x \in \mathcal{X}^+$. In words, the forecaster is well-calibrated if among all those days for which the prediction is x , the proportion of rainy days is also x , and this is true for every value of x . In meteorology, the criterion of calibration is referred to as validity (Miller, 1962), or reliability (Murphy, 1973), and the well-calibrated forecaster is said to be perfectly reliable.

For obvious reasons, being well-calibrated is usually regarded as a desirable characteristic of a forecaster. It has been pointed out elsewhere (DeGroot, 1979), however, that it is typically easy for any forecaster to make himself well-calibrated by specifying predictions that do not represent his subjective probabilities and in which he does not believe. Furthermore, as Dawid (1980) has stated, even if the forecaster's true probabilities make him well-calibrated, "this does not necessarily mean that they are 'accurate' in all respects; and even if they are accurate, they may not be of much substantive value if the forecaster is a poor meteorologist." Thus, a well-calibrated forecaster is not necessarily a good forecaster, and we shall now consider the problem of comparing well-calibrated forecasters.

3. Refinement

In this section, we shall restrict attention to well-calibrated forecasters. Let μ denote the relative frequency of days on which it rains. In



meteorology, μ is sometimes called the climatological probability. For any well-calibrated forecaster, it must be true that

$$\sum_{x \in \mathcal{X}} x\nu(x) = \mu \ . \tag{3.1}$$

Throughout this paper we shall assume that $0 < \mu < 1$.

In order to emphasize the possible differences that can exist among such forecasters, we shall begin by considering two extreme types. Suppose that $\mu \in \mathcal{X}$. Then the forecaster A_0 whose prediction each day is μ will be well-calibrated, although his predictions are completely useless for any purpose whatsoever. The predictions of A_0 are characterized by the degenerate probability function

$$\begin{aligned} \nu_{A_0}(\mu) &= 1 \ , \\ \nu_{A_0}(x) &= 0 \quad \text{for} \quad x \neq \mu \ . \end{aligned} \tag{3.2}$$

When $\mu \in \mathcal{X}$, we shall refer to A_0 as the least-refined forecaster.

Next, consider a well-calibrated forecaster A^0 whose predictions are characterized by the following probability function:

$$\begin{aligned} \nu_{A^0}(1) &= \mu \ , \\ \nu_{A^0}(0) &= 1-\mu \ , \\ \nu_{A^0}(x) &= 0 \text{ for } x \neq 0,1 \ . \end{aligned} \tag{3.3}$$

It can be seen from (3.3) that the only probabilities of rain that forecaster A^0 ever specifies are 0 and 1 , and since A^0 is well-calibrated, his predictions are always correct. We shall refer to A^0 as the most-refined forecaster. In meteorology, A_0 is said to exhibit zero sharpness, and A^0 to exhibit perfect sharpness (Sanders, 1963).

It is clear from A_0 and A^0 that quite different types of behavior are possible among well-calibrated forecasters, ranging from useless to perfect predictions. We shall now describe a concept that yields a partial ordering on the class of all well-calibrated forecasters and justifies our referring to A_0 and A^0 as the least and the most refined members of this class.

A <u>stochastic transformation</u> $h(x|y)$ is a function defined on $\mathcal{X}\times\mathcal{X}$ such that

$$\begin{aligned} &h(x|y) \geq 0 \text{ for } x\in\mathcal{X} \text{ and } y\in\mathcal{X} , \\ &\sum_{x\in\mathcal{X}} h(x|y) = 1 \text{ for } y\in\mathcal{X} . \end{aligned} \tag{3.4}$$

Now consider two arbitrary well-calibrated forecasters whose predictions are characterized by the probability functions $\nu_A(x)$ and $\nu_B(x)$. We say that A is <u>at least as refined as</u> B if there exists a stochastic transformation h such that the following relations are satisfied:

$$\sum_{y\in\mathcal{X}} h(x|y)\nu_A(y) = \nu_B(x) \quad \text{for } x\in\mathcal{X} , \tag{3.5}$$

$$\sum_{y\in\mathcal{X}} h(x|y)y\nu_A(y) = x\nu_B(x) \quad \text{for } x\in\mathcal{X} . \tag{3.6}$$

By subtracting (3.6) from (3.5) we get

$$\sum_{y\in\mathcal{X}} h(x|y)(1-y)\ \nu_A(y) = (1-x)\ \nu_B(x) \quad \text{for } x\in\mathcal{X} , \tag{3.7}$$

which adds a touch of symmetry when (3.7) is paired with (3.6).

Together, the relations (3.5) and (3.6), or (3.6) and (3.7), state that if we know the predictions of forecaster A, then we can simulate the predictions of forecaster B by using an auxiliary randomization based on the

stochastic transformation h as follows: If A makes the prediction y on a particular day, then we simulate a prediction x in accordance with the conditional probability distribution $h(x|y)$. The prediction x will then have exactly the same probabilistic properties as the predictions of forecaster B. The relation (3.5) guarantees that we will obtain each prediction x with the same frequency $\nu_B(x)$ that B does, and the relation (3.6) guarantees that our predictions will still be well-calibrated.

To see that any forecaster A is at least as refined as the least-refined forecaster A_0, let us define the stochastic transformation h as follows:

$$\begin{aligned} h(\mu|y) &= 1 \quad \text{for } y\in\mathcal{X}, \\ h(x|y) &= 0 \quad \text{for } x \neq \mu . \end{aligned} \tag{3.8}$$

Then it follows from (3.1) that (3.5) and (3.6) are satisfied when ν_B is replaced by ν_{A_0} as defined by (3.2).

Similarly, to see that the most refined forecaster A^0 is at least as refined as any other forecaster B, let us define the stochastic transformation h as follows.

$$\begin{aligned} h(x|1) &= \frac{1}{\mu}\, x\nu_B(x) \quad \text{for } x\in\mathcal{X}, \\ h(x|0) &= \frac{1}{1-\mu}\,(1-x)\nu_B(x) \quad \text{for } x\in\mathcal{X} . \end{aligned} \tag{3.9}$$

Since B is well-calibrated, it follows from (3.1) that the function h defined in (3.9) has the properties required of a stochastic transformation. The definition of $h(x|y)$ for $y \neq 0,1$ is irrelevant since forecaster A^0 never makes a prediction other than 0 or 1. The relations (3.5) and (3.6) will now be satisfied when ν_A is replaced by ν_{A^0} as defined by (3.3).

Since the relationship among well-calibrated forecasters defined by the concept of one being at least as refined as another is both reflexive and transitive, this relationship induces a partial ordering among those forecasters. We do not obtain a total ordering, however, as the next example shows.

Suppose that A and B are well-calibrated forecasters characterized by the following probability functions:

$$\nu_A(x) = \begin{cases} .1 & \text{for } x = 0, \\ .8 & \text{for } x = .5, \\ .1 & \text{for } x = 1, \end{cases} \tag{3.10}$$

and

$$\nu_B(x) = \begin{cases} .5 & \text{for } x = .1, \\ .5 & \text{for } x = .9. \end{cases} \tag{3.11}$$

Here $\mu = .5$.

In this example, A is not at least as refined as B. To see this, suppose on the contrary that there were a stochastic transformation $h(x|y)$ that satisfied (3.5) and (3.6), and let

$$a = h(.1|0), \quad b = h(.1|.5), \quad \text{and} \quad c = h(.1|1). \tag{3.12}$$

Then for $x = .1$, the relations (3.5) and (3.6) become

$$\begin{aligned} (.1)a + (.8)b + (.1)c &= .5, \\ (.4)b + (.1)c &= .05. \end{aligned} \tag{3.13}$$

The two equations in (3.13) imply that $a-c=4$, which is an impossibility since both $0 \leq a \leq 1$ and $0 \leq c \leq 1$.

On the other hand, neither is B at least as refined as A. To see this, we need only note that on 20 percent of the days, A makes predictions of rain or no rain that are certain to be correct (because A is well-calibrated)

whereas B never makes correct predictions with certainty. Thus, in this example neither A nor B is at least as refined as the other.

4. Sufficiency

In Section 2 we characterized the predictive behavior of any forecaster, regardless of whether or not he was well-calibrated, by the functions $v(x)$ and $\rho(x)$. In effect, we represented the joint distribution of the prediction x and the indicator of rain in terms of the marginal distribution of x and the conditional probability $\rho(x)$ of rain given the prediction x. But it is also useful at times to use an alternative factorization of this joint distribution (see, e.g., Lindley, Tversky, and Brown, 1979, and Lindley, 1981).

Let θ denote the indicator of rain, so $\theta = 1$ if rain occurs on a particular day and $\theta = 0$ otherwise, and for any given forecaster let $f(x|\theta)$ denote the conditional probability function of the forecaster's predictions given θ. In other words, for $\theta = 1$, $f(x|\theta)$ represents the frequency function of the forecaster's predictions on days when rain actually occurs. It follows that for $x \in \mathcal{X}$,

$$\mu f(x|1) = \rho(x)v(x) , \tag{4.1}$$

$$(1-\mu)f(x|0) = [1-\rho(x)]v(x) . \tag{4.2}$$

It follows from (4.1) that the probability functions $f(x|\theta)$ for $\theta=0$ and $\theta=1$ characterize the forecaster's predictive behavior.

Now consider two forecasters A and B characterized by the functions $f_A(x|\theta)$ and $f_B(x|\theta)$. Following the original work of Blackwell (1951, 1953) on the comparison of experiments, we say that forecaster A is <u>sufficient</u> for forecaster B if there exists a stochastic transformation $h(x|y)$ such that

$$\sum_{y\in\mathcal{X}} h(x|y) f_A(y|\theta) = f_B(x|\theta) \quad \text{for} \quad x\in\mathcal{X} \quad \text{and} \quad \theta=0,1 \tag{4.3}$$

(see, e.g., DeGroot, 1970, Sec. 14.17). The interpretation of (4.3) is similar to that given in Section 3: forecaster A is sufficient for forecaster B if we can simulate the predictions of B from the predictions of A by using an auxiliary randomization based on the stochastic transformation h .

As before, the relationship of sufficiency induces a partial ordering among all forecasters. Since we have applied this relationship to all forecasters, however, and not just to those who are well-calibrated, it is not necessarily true that if A is sufficient for B then A is at least as "good" a forecaster as B . For example, suppose that A never makes a prediction other than $x=0$ or $x=1$, but that he is always wrong about whether or not it is going to rain. Then A is sufficient for every other forecaster, even though he is the worst possible forecaster. Of course, if we knew that A was always wrong, his predictions would be just as useful to us as those of a forecaster who was always correct.

Theorem 1. Consider two forecasters A and B whose predictive behavior is characterized by the functions $\nu_A(x)$, $\rho_A(x)$, $\nu_B(x)$, and $\rho_B(x)$. Then forecaster A is sufficient for forecaster B if and only if there exists a stochastic transformation h such that the following relations are satisfied:

$$\sum_{y\in\mathcal{X}} h(x|y)\nu_A(y) = \nu_B(x) \quad \text{for} \quad x\in\mathcal{X} \ , \tag{4.4}$$

$$\sum_{y\in\mathcal{X}} h(x|y)\rho_A(y)\nu_A(y) = \rho_B(x)\nu_B(x) \quad \text{for} \quad x\in\mathcal{X} \ . \tag{4.5}$$

Proof. Consider any fixed value $x \in \mathcal{X}$. It follows from (4.1) that for $\theta=1$, the relation (4.3) is the same as (4.5). Furthermore, it follows from (4.2) that for $\theta=0$, the relation (4.3) is the same as the relation

$$\sum_{y \in \mathcal{X}} h(x|y)[1-\rho_A(y)]\nu_A(y) = [1-\rho_B(x)]\nu_B(x) , \qquad (4.6)$$

which, in view of (4.5), is equivalent to (4.4). ■

Recall now that a forecaster is well-calibrated if $\rho_A(x) = x$ for all $x \in \mathcal{X}^+$. The following result follows immediately in the light of relations (3.5) and (3.6).

Theorem 2. Consider two well-calibrated forecasters A and B. Then forecaster A is at least as refined as forecaster B if and only if forecaster A is sufficient for forecaster B.

5. Conditions for sufficiency

In this section we shall again consider two well-calibrated forecasters A and B. In order to determine whether or not A is sufficient for B based on the discussion in the previous sections, it is necessary to determine whether or not there exists a stochastic transformation that satisfies either the relations (3.5) and (3.6) or the relations (4.3). Attempts to establish the existence or non-existence of such a stochastic transformation can be frustrating and fruitless. Fortunately, Blackwell and Girshick (1954) and Bradt and Karlin (1956) have provided some direct methods for determining whether or not A is sufficient for B that eliminate the necessity of having to consider stochastic transformations.

For any forecaster, A , let

$$\alpha_A(x) = f_A(x|1) + f_A(x|0) \quad \text{for} \quad x\epsilon\mathcal{X} , \tag{5.1}$$

and for $0 \leq t \leq 1$, let $\mathcal{A}_A(t)$ denote the subset of points in $\mathcal{X}$ such that $f_A(x|1) < t\,\alpha_A(x)$. Furthermore, let

$$F_A(t) = \sum_{x\epsilon\mathcal{A}_A(t)} \alpha_A(x) \quad \text{for} \quad 0 \leq t \leq 1 \tag{5.2}$$

and

$$C_A(t) = \int_0^t F_A(u)\,du \quad \text{for} \quad 0 \leq t \leq 1 . \tag{5.3}$$

Then, as demonstrated in Theorem 12.4.1 of Blickwell and Girshick (1954), forecaster A is sufficient for forecaster B if and only if $C_A(t) \geq C_B(t)$ for all t in the interval $0 \leq t \leq 1$.

A brief heuristic interpretation of this result is as follows. Suppose that the parameter θ has prior probabilities given by $Pr(\theta=1)=Pr(\theta=0)=\frac{1}{2}$. Then $\frac{1}{2}\alpha_A(x)$ is the marginal distribution of x for forecaster A . Furthermore, if we let $\pi_A(x)$ denote the posterior probability $Pr(\theta=1|x)$ for forecaster A , then $\mathcal{A}_A(t)$ denotes the set of values of x for which $\pi_A(x) < t$. It can now be seen from (5.2) that $\frac{1}{2}F_A(t)$ is the distribution function of the posterior probability $\pi_A(x)$ for forecaster A . For an informative forecaster, the values of $\pi_A(x)$ will tend to be concentrated near 0 and 1 , and away from their mean value $E_A[\pi_A(x)] = \frac{1}{2}$. The condition that $C_A(t) \geq C_B(t)$ for all t in the interval $0 \leq t \leq 1$ is equivalent to the condition that $E_A\{\varphi[\pi_A(x)]\} \geq E_B\{\varphi[\pi_B(x)]\}$ for every continuous convex function φ . In this sense, the condition expresses the notion that the probability distribution of $\pi_A(x)$ is more spread out from $\frac{1}{2}$ than the probability distribution of $\pi_B(x)$.

We are now ready to establish the main result of this section. Recall that the set $\mathcal{X}$ comprises the points $x_0 < x_1 < \ldots < x_k$.

Theorem 3. Consider two well-calibrated forecasters A and B. Then forecaster A is sufficient for forecaster B if and only if the following inequalities are satisfied:

$$\sum_{i=0}^{j-1} (x_j - x_i)[\nu_A(x_i) - \nu_B(x_i)] \geq 0 \quad \text{for } j=1,\ldots, k-1 . \tag{5.4}$$

Proof. Since both A and B are well-calibrated, it follows from (4.1) and (4.2) that

$$\frac{f_A(x|1)}{\alpha_A(x)} = \frac{f_B(x|1)}{\alpha_B(x)} = \frac{(x/\mu)}{(x/\mu) + [(1-x)/(1-\mu)]} \tag{5.5}$$

whenever $\alpha_A(x)$ and $\alpha_B(x)$ are non-zero. Even if either $\alpha_A(x)$ or $\alpha_B(x)$ is zero for some $x\in\mathcal{X}$, without loss of generality we still may define (5.5) to be satisfied. Next, for $0 \leq t \leq 1$ and $x\in\mathcal{X}$, define

$$s(t,x) = t\left(\frac{1-x}{1-\mu}\right) - (1-t)\frac{x}{\mu} . \tag{5.6}$$

Then both the sets $\mathcal{A}_A(t)$ and $\mathcal{A}_B(t)$ contain precisely those points $x\in\mathcal{X}$ for which $s(t,x)>0$. Since the sets $\mathcal{A}_A(t)$ and $\mathcal{A}_B(t)$ are identical, we shall denote this common set simply by $\mathcal{A}(t)$.

From (5.2) and (5.3) we can write

$$C_A(t) = \int_0^t [\sum_{x\in\mathcal{A}(u)} \alpha_A(x)]du .$$

For each $x\in\mathcal{X}$, $\alpha_A(x)$ contributes to the integral over a certain set of u-values of length $t-f(x|1)/\alpha_A(x)$. Thus we can re-express $C_A(t)$ as

$$C_A(t) = \sum_{x\in\mathcal{A}(t)} [t\alpha_A(x) - f_A(x|1)] . \tag{5.7}$$

Next, using (4.1) and (4.2) and the fact that A is well-calibrated. we can rewrite (5.7) as follows:

$$C_A(t) = \sum_{x \in \mathcal{S}(t)} s(t,x)\nu_A(x) \ . \tag{5.8}$$

Furthermore, if we rewrite $s(t,x)$, as given by (5.6), in the form

$$s(t,x) = \frac{1}{\mu(1-\mu)} \{t\mu - [t\mu+(1-t)(1-\mu)]x\} \ , \tag{5.9}$$

then it can be seen that $\mathcal{S}(t)$ contains precisely these points $x \in \mathcal{X}$ for which the quantity inside braces in (5.9) is positive. Thus, $C_A(t)$ can be expressed as follows:

$$C_A(t) = \frac{1}{\mu(1-\mu)} \sum_{x \in \mathcal{X}} \{t\mu-[t\mu+(1-t)(1-\mu)]x\}^+ \nu_A(x) \ , \tag{5.10}$$

where, as usual, the notation $(m)^+$ denotes the positive part of the quantity m.

For forecaster B, the function $C_B(t)$ is also given by (5.10) with $\nu_A(x)$ replaced by $\nu_B(x)$. Let

$$L_A(t) = \mu(1-\mu)\ C_A(t) \tag{5.11}$$

and let $L_B(t)$ be defined similarly. Then it follows from Theorem 12.4.1 of Blackwell and Girshick (1954), as cited earlier, that forecaster A is sufficient for forecaster B if and only if $L_A(t) \geq L_B(t)$ for all t in the interval $0 \leq t \leq 1$.

Corresponding to the points $0 \leq x_0 < x_1 < \ldots < x_k \leq 1$ in $\mathcal{X}$, let the points $0 \leq t_0 < t_1 < \ldots < t_k \leq 1$ be defined by the relations

$$t_j\mu - [t_j\mu+(1-t_j)(1-\mu)]x_j = 0 \quad \text{for} \quad j = 0,1,\ldots, k \ . \tag{5.12}$$

Then both $L_A(t)$ and $L_B(t)$ are continuous, piecewise linear functions over the interval $0 \leq t \leq 1$ with $L_A(0) = L_B(0) = 0$ and $L_A(1) = L_B(1) = 1$, and

with vertices at the points $t_0, t_1, \ldots, t_k$. Furthermore, $L_A(t_0) = L_B(t_0) = 0$ and, for $j = 1, \ldots, k$,

$$\frac{L_A(t_j)}{t_j\mu+(1-t_j)(1-\mu)} = \sum_{i=0}^{j-1} (x_j - x_i)\nu_A(x_i) , \tag{5.13}$$

with an analogous expression for forecaster B.

Finally, we note that when $j=k$, the right-hand side of (5.13) can be reduced as follows by using (3.1):

$$\begin{aligned} \sum_{i=0}^{k-1} (x_k - x_i)\nu_A(x_i) &= x_k[1-\nu_A(x_k)] - [\mu - x_k\nu_A(x_k)] \\ &= x_k - \mu . \end{aligned} \tag{5.14}$$

Thus, $L_A(t_k) = L_B(t_k)$. We have now established that forecaster A is sufficient for forecaster B if and only if $L_A(t_j) \geq L_B(t_j)$ for $j=1, \ldots, k-1$. It follows from (5.13) that these $k-1$ inequalities are equivalent to the $k-1$ inequalities (5.4). ■

We make use of this theorem in the next section.

6. The least-refined, well-calibrated forecaster

In Section 3, we required that $\mu \in \mathcal{X}$ in order to ensure that the least-refined forecaster who is well-calibrated will always announce μ as his forecast. Suppose now that $\mu \notin \mathcal{X}$, and let x_L and x_U be the pair of adjacent values in $\mathcal{X}$ just bracketing μ, i.e., $x_L < \mu < x_U$. Then let A_0' be a well-calibrated forecaster characterized by the probability function:

$$\begin{aligned} \nu_{A_0'}(x_U) &= (\mu - x_L)/(x_U - x_L) , \\ \nu_{A_0'}(x_L) &= (x_U - \mu)/(x_U - x_L) , \\ \nu_{A_0'}(x) &= 0 \text{ for } x \neq x_U \text{ or } x_L . \end{aligned} \tag{6.1}$$

It can be seen from (6.1) that A_0' concentrates his forecasts as closely as possible to μ given the permissible forecast values. Thus, there is an intuitive sense in which A_0' is not as refined as another forecaster who spreads out his probabilities over at least some of the other values of x. We make this notion precise in the following theorem.

Theorem 4. The well-calibrated forecaster A_0', whose probability function $\nu_{A_0'}(x)$ is given by (6.1), is least refined among all other well-calibrated forecasters.

Proof. Consider any other well-calibrated forecaster A. Then, from Theorem 2, A is at least as refined as A_0' if and only if A is sufficient for A_0', and, from Theorem 3, this is true if and only if

$$\sum_{i=0}^{j-1} (x_j - x_i)[\nu_A(x_i) - \nu_{A_0'}(x_i)] \geq 0 \quad \text{for} \quad j=1,2,\ldots, k-1 . \tag{6.2}$$

To verify (6.2), we note that for $j=1,\ldots, L$,

$$\sum_{i=0}^{j-1} (x_j - x_i)[\nu_A(x_i) - \nu_{A_0'}(x_i)] = \sum_{i=0}^{j-1} (x_j - x_i)\nu_A(x_i) , \tag{6.3}$$

which clearly is nonnegative since $x_j > x_i$ and $\nu_A(x)$ is a probability function. For $j=U$, recalling that $U = L+1$, we have

$$\begin{aligned} &\sum_{i=0}^{L} (x_U - x_i)[\nu_A(x_i) - \nu_{A_0'}(x_i)] \\ &= \sum_{i=0}^{L} (x_U - x_i)\nu_A(x_i) - (x_U - x_L)\nu_{A_0'}(x_L) \\ &= \sum_{i=0}^{L} (x_U - x_i)\nu_A(x_i) - (x_U - \mu) , \end{aligned} \tag{6.4}$$

where the final expression follows from (6.1). Since A is well-calibrated, we can now use (3.1) to rewrite the expression following the final equality sign in (6.4) as follows:

$$\sum_{i=0}^{L}(x_U-x_i)\nu_A(x_i) - \sum_{i=0}^{k}(x_U-x_i)\nu_A(x_i)$$

$$= \sum_{i=U}^{k}(x_i-x_U)\nu_A(x_i) \geq 0 \ . \tag{6.5}$$

Similarly, for $j=U+1,\ldots, k$, the left-hand side of expression (6.2) equals

$$\sum_{i=j}^{k}(x_i-x_j)\nu_A(x_i) \geq 0 \ . \ \blacksquare$$

The use of Theorem 3 is critical to the preceding proof, for otherwise we would need to construct the actual stochastic transformation $h(x|y)$ going from ν_A to $\nu_{A'_o}$, simultaneously ensuring that the calibration condition holds. We have found this to be a nontrivial task.

7. Scoring rules for assessment

In the television station example introduced in Section 1, we get to see a finite set of forecasts and the associated indicators of whether or not rain occurred, i.e., $\{(p_j, y_j): j = 1,2,\ldots, n\}$. Several authors have suggested scoring rules to be used to assess probability assessors in such situations. Here we relate some of these to the probabilistic concepts of calibration and refinement.

One of the earliest scoring rule proposals suggested in the context of meteorological forecasts is the "Brier Score",

$$BS_n = \frac{1}{n} \sum_{j=1}^{n} (p_j - y_j)^2 , \tag{7.1}$$

which the forecaster is to attempt to minimize (Brier, 1950). In the case of binary outcomes (rain, no rain), Winkler (1967) notes that the Brier Score is equivalent to the general quadratic scoring rule proposed by de Finetti (1965), designed to oblige the forecaster "to express his true feelings" (de Finetti, 1962).

Other general classes of "strictly proper" scoring rules include Good's (1952) logarithmic scoring rule and the spherical scoring rule (see Staël von Holstein 1970, and Savage 1971).

If we let n_i equal the number of days out of n on which the forecaster predicts rain with probability x_i, and r_i the number of these n_i days on which it actually does rain, we can rewrite the Brier Score of (7.1) as

$$BS_n = \frac{1}{n} \sum_{i=0}^{k} n_i (x_i - \frac{r_i}{n_i})^2 + \frac{1}{n} \sum_{i=0}^{k} n_i \frac{r_i}{n_i} (1 - \frac{r_i}{n_i}) , \tag{7.2}$$

or as

$$BS_n = \frac{1}{n} \sum_{i=0}^{k} n_i (x_i - \frac{r_i}{n_i})^2 + \frac{r}{n} (1 - \frac{r}{n}) - \frac{1}{n} \sum_{i=0}^{k} n_i (\frac{r_i}{n_i} - \frac{r}{n})^2 , \tag{7.3}$$

where $\sum_{i=0}^{k} r_i = r$. Tukey, Mosteller, and Fienberg (1965) suggest a variant of (7.2) which essentially allows the two components on the right-hand side to be given different weights.

To understand how the components of the Brier Score relate to the concepts discussed here we let $n \to \infty$ in such a manner that $r_i/n_i \to \rho(x_i)$ and $n_i/n \to \nu(x_i)$. Then

$$\lim_{n\to\infty} \frac{r_i}{n} = \rho(x_i)\nu(x_i) \tag{7.4}$$

and

$$\lim_{n\to\infty} \frac{r_i^2}{n_i n} = \rho^2(x_i)\nu(x_i) \ . \tag{7.5}$$

Any sampling scheme of trials with these limiting properties suffices for our purposes. Thus, from (7.2) we have

$$\begin{aligned} BS &= \lim_{n\to\infty} BS_n \\ &= \sum_{i=0}^{k} \nu(x_i)\,[x_i - \rho(x_i)]^2 + \sum_{i=0}^{k} \nu(x_i)\rho(x_i)[1 - \rho(x_i)] \ . \end{aligned} \tag{7.6}$$

The first term on the right-hand side of (7.6) is the weighted mean square difference between the forecasted probability x_i and the frequency of rain $\rho(x_i)$. As such it is a measure of calibration. If the forecaster is well-calibrated, this term equals zero.

The second term on the right-hand side of (7.6) measures the dispersion of the results of the forecaster's predictions. As such it rewards the forecaster for spreading his predictions as much as possible, and thus is a measure of the forecaster's refinement. The following theorem shows that there is a direct relationship between this term and the concepts of refinement and sufficiency presented in Sections 3 and 4.

Theorem 5. If forecaster A is sufficient for forecaster B , then

$$\sum_{x\in\mathcal{X}} \nu_A(x)\rho_A(x)[1 - \rho_A(x)] \leq \sum_{x\in\mathcal{X}} \nu_B(x)\rho_B(x)[1 - \rho_B(x)] \ . \tag{7.7}$$

Proof: Since A is sufficient for B , from (4.5) we have

$$\sum_x \rho_B(x)\nu_B(x) = \sum_x \sum_y h(x|y)\rho_A(y)\nu_A(y)$$

$$= \sum_y [\sum_x h(x|y)]\rho_A(y)\nu_A(y) \qquad (7.8)$$

$$= \sum_y \rho_A(y)\nu_A(y) \ .$$

Next, by applying both (4.4) and (4.5) we have

$$\nu_B(x)\rho_B^2(x) = \frac{[\sum_y h(x|y)\rho_A(y)\nu_A(y)]^2}{\sum_y h(x|y)\nu_A(y)}$$

$$= [\sum_y h(x|y)\nu_A(y)] \, [\sum_y \frac{h(x|y)\nu_A(y)}{\sum_{y'} h(x|y')\nu_A(y')} \rho_A(y)]^2$$

$$\leq [\sum_y h(x|y)\nu_A(y)] \, [\sum_y \frac{h(x|y)\nu_A(y)}{\sum_{y'} h(x|y')\nu_A(y')} \rho_A^2(y)]$$

$$= \sum_y h(x|y)\nu_A(y)\rho_A^2(y) \ , \qquad (7.9)$$

where the inequality is a special case of Jensen's inequality. Summing (7.9) over x now yields

$$\sum_x \nu_B(x)\rho_B^2(x) \leq \sum_x \sum_y h(x|y)\nu_A(y)\rho_A^2(y)$$

$$= \sum_y [\Sigma h(x|y)]\nu_A(y)\rho_A^2(y)$$

$$= \sum_y \nu_A(y)\rho_A^2(y) \ . \qquad (7.10)$$

Finally, combining (7.8) and (7.10) yields the inequality (7.7). ■

We recall from Section 5 that forecaster A is sufficient for forecaster B if and only if $C_A(t) \geq C_B(t)$, where $C_A(t)$ is defined by (5.3), and that this condition is equivalent to

$$E_A\{\varphi[\pi_A(x)]\} \geq E_B\{\varphi[\pi_B(x)]\} \tag{7.11}$$

for every continuous convex function φ. Theorem 5 is, in effect, a special case of this equivalence. From (7.11) we can construct a class of generalized limiting scoring rules that replace the second term of (7.6) by

$$\sum_{x\in\mathcal{X}} \nu(x)\varphi[\rho(x)]. \tag{7.12}$$

The actual assessment of probability assessors based on a finite set of forecasts requires a careful description of the stochastic mechanisms associated with the production of predictions for the forecasters being compared. We shall present such a description in a separate paper.

8. Multivariate forecasts

In the preceding sections we have considered events with $s = 2$ possible outcomes (e.g., rain, no rain). Yet climatological forecasting often involves $s > 2$ outcomes (e.g., rain, snow, and neither rain nor snow, or a set of temperature ranges). In such situations the probability assessor specifies a vector of probabilities $\underset{\sim}{x}$, restricted to a finite set of values lying in the $(s-1)$-dimensional simplex. If the conditional probabilities of the s outcomes given the prediction $\underset{\sim}{x}$ is represented in vector form by $\underset{\sim}{\rho}(\underset{\sim}{x})$, then the multivariate forecaster is well-calibrated if $\underset{\sim}{\rho}(\underset{\sim}{x}) = \underset{\sim}{x}$ for all $\underset{\sim}{x}\in\mathcal{X}$. Note that this well-calibrated multivariate forecaster is also

well-calibrated, in the sense of Section 2, for each binary problem formed by combining the s outcomes into two groups; however, a forecaster who is "marginally" well-calibrated for predicting "rain" or "no rain" may no longer be well-calibrated when "no rain" is divided into two or more possible outcomes.

More formally, let $\underset{\sim}{x} = (x_1, \ldots, x_s)$ and $\underset{\sim}{\rho}(\underset{\sim}{x}) = [\rho_1(\underset{\sim}{x}), \ldots, \rho_s(\underset{\sim}{x})]$. Furthermore, let $\mathcal{J} = \{I_1, \ldots, I_k\}$ represent a partition of the set $\{1, \ldots, s\}$ into k nonempty, mutually exclusive, and exhaustive sets $I_1, \ldots, I_k$. Then a forecaster is said to be <u>marginally well-calibrated with respect to the partition</u> $\mathcal{J}$ if

$$\sum_{i \in I_j} \rho_i(\underset{\sim}{x}) = \sum_{i \in I_j} x_i \qquad \text{for } j = 1, \ldots, k \text{ and } \underset{\sim}{x} \in \mathcal{X}. \tag{8.1}$$

Similarly, we can develop the notion of conditionally well-calibrated forecasters. Consider again the problem treated in Sections 2-7, in which $s = 2$ and the forecaster simply specifies his probability x of rain. The forecaster may be well-calibrated for some, but not all, values of x. In other words, it may be true that $\rho(x) = x$ when x belongs to some subset $\mathcal{X}_0$ of $\mathcal{X}$, but not for all values of $x \in \mathcal{X}$. In this case, we may say that the forecaster is conditionally well-calibrated, given that $x \in \mathcal{X}_0$.

Now consider the general multivariate forecasting problem introduced in this section. Let the partition $\mathcal{J}$ be as defined here, and let $\mathcal{X}_0$ denote a proper subset of $\mathcal{X}$. Then a forecaster is said to be conditionally well-calibrated with respect to the partition $\mathcal{J}$, given that $\underset{\sim}{x} \in \mathcal{X}_0$, if the relation (8.1) is satisfied for $j = 1, \ldots, k$ and all $\underset{\sim}{x} \in \mathcal{X}_0$.

For well-calibrated multivariate forecasters, we can define the concept of refinement by means of a multivariate stochastic transformation. Moreover, this notion of refinement can again be directly linked to sufficiency in the comparison of experiments with a finite number of outcomes. Finally, the concept of one forecaster being marginally or conditionally more refined than another can be developed.

Critical to the multivariate versions of calibration and refinement as proposed in this section is the orientation of the vector of forecasted probabilities $\underset{\sim}{x}$. Each component of $\underset{\sim}{x}$ refers to a specific outcome. This methodology should be contrasted with the multivariate approach, described for example by Lichtenstein, Fischhoff, and Phillips (1977), in which the forecaster "selects the single most likely alternative and states the probability that it is correct." Kadane and Lichtenstein (1981) show that such a loss of orientation leads to the inability to recalibrate a forecaster's assessments. From the discussion here, it should be clear that a careful description of calibration and refinement in both the binary and multivariate settings requires a well-specified set of outcomes, and probability assessments specifically tied to those outcomes.

References

Blackwell, D. (1951). Comparison of experiments. Proc. Second Berkeley Symp. Math. Statist. Probability. Berkeley: University of California Press. 93-102.

Blackwell, D. (1953). Equivalent comparison of experiments. Ann. Math. Statist. 24 265-272.

Blackwell, D. and Girshick, M.A. (1954). Theory of Games and Statistical Decisions. New York: John Wiley.

Bradt, R.N. and Karlin S. (1956). On the design and comparison of certain dichotomous experiments Ann. Math. Statist. 27 390-409.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1-3.

Dawid, A.P. (1980). Discussion of papers on "Improving judgements using feedback." In J.M. Bernardo et al., eds., Bayesian Statistics, 418-419. Proc. of the First Int. Meeting. Valencia, Spain: University Press

Dawid, A.P. (1981). The well-calibrated Bayesian. J. Amer. Statist. Assoc. (forthcoming).

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. (English translation from French). In H.E. Kyburg and H.E. Smokler, eds., Studies in Subjective Probability (1964). New York: John Wiley, 93-158.

de Finetti, B. (1962). Does it make sense to speak of 'Good Probability Appriasers'?. In I. J. Good, gen. ed., The Scientist Speculates--An Anthology of Partly-Baked Ideas. New York: Basic Books, 357-63.

de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. British J. Mathematical and Statistical Psychology, 18 87-123.

DeGroot, M.H. (1970). Optimal Statistical Decisions, New York: McGraw-Hill.

DeGroot, M.H. (1979). Comments on Lindley, et al. J. Roy. Statist. Soc. (A) 142, 172-173.

Good, I.J. (1952). Rational decisions. J. Roy. Statist. Soc. (B). 14, 107-114.

Kadane, J.B., and Lichtenstein, S. (1981). Calibration in perspective. Unpublished manuscript.

Lichtenstein, S., Fischhoff, B. and Phillips, L.D. (1977). Calibration of probabilities: the state of the art. In H. Jungermann and G. de Zeeuw, eds., Decision Making and Change in Human Affairs. Dordrecht, Holland: D. Reidel Publishing Company, 275-324.

Lindley, D.V. (1981). The improvement of probability judgements. Unpublished manuscript.

Lindley, D.V., Tversky, A., and Brown, R.V. (1979). On the reconciliation of probability assessments. J. Roy. Statist. Soc. (A) 142, 146-180.

Miller, R.G. (1962). Statistical prediction by discriminant analysis. Meteorological Monographs. 4 No. 25.

Murphy A. H. (1973). A new vector partition of the probability score. J. Applied Meteorology 12, 595-600.

Murphy, A.H. and Winkler, R.L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. Applied Statistics 26, 41-47.

Pratt, J.W. (1962). Must subjective probabilities be realized as relative frequencies? Unpublished seminar paper. Harvard University Grad. School of Bus. Administration.

Sanders, F. (1963). On subjective probability forecasting. J. Applied Meteorology, 2, 191-201.

Savage, L.J. (1971). Elicitation of personal probabilities and expectations. J. Amer. Statist. Assoc. 66, 783-801.

Staël von Holstein, C.-A. S. (1970). Assessment and Evaluation of Subjective Probability Distributions Stockholm: Economic Research Institute, Stockholm School of Economics.

Tukey, J.W., Mosteller, F. and Fienberg, S.E. (1965). Scoring probability forecasts. Memorandum NS-37, Dept. of Statistics, Harvard University.

Winkler, R.L. (1967). The quantification of judgment: some methodological suggestions. J. Amer. Statist. Assoc. 62, 1105-1120.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report #205	2. GOVT ACCESSION NO. AD-A104 174	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ASSESSING PROBABILITY ASSESSORS: CALIBRATION AND REFINEMENT		5. TYPE OF REPORT & PERIOD COVERED To May, 1981
		6. PERFORMING ORG. REPORT NUMBER T.R. #205
7. AUTHOR(s) Morris H. DeGroot Stephen E. Fienberg		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Carnegie-Mellon University Pittsburgh, PA 15213		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Contracts Office Carnegie-Mellon University Pittsburgh, PA 15213		12. REPORT DATE 13. NUMBER OF PAGES 27
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release: Distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

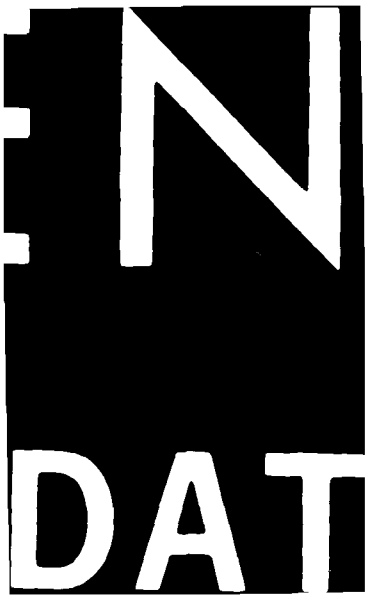
18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


N
DAT